METADATA: Small Area Indicators File

Dataset: small_area_indicators.csv

OVERVIEW

- **Purpose**: Small area estimation inputs and outputs for poverty mapping
- **Geographic Level**: Sub-district administrative units
- **Methodology**: Fay-Herriot and EBLUP models
- **Coverage**: 500 small areas across 150 districts
- **Reference Period**: 2024 estimates based on 2022 census and 2024 survey

SMALL AREA ESTIMATION FRAMEWORK

- **Direct Estimator**: Horvitz-Thompson from survey data
- **Model-Based**: Fay-Herriot area-level model
- **Auxiliary Data**: Census and administrative sources
- **Validation**: Benchmarked to district estimates

VARIABLE DEFINITIONS

IDENTIFICATION

- **area_id** [String]: Unique small area identifier (SA####)
- **district_code** [String]: Parent district code (D###)
- **area_name** [String]: Small area name
- **area_type** [String]: Area classification
  - Urban formal: Planned urban areas
  - Urban informal: Informal settlements
  - Rural village: Traditional villages
  - Rural scattered: Dispersed settlements
  - Commercial: Business districts
  - Industrial: Manufacturing areas

POPULATION CHARACTERISTICS

- **total_population** [Integer]: Estimated population

- **sample_size** [Integer]: Survey sample size in area

DIRECT ESTIMATES

- **direct_estimate_poverty** [Numeric]: Direct poverty rate (0-1)
- **direct_estimate_se** [Numeric]: Standard error of direct estimate

AUXILIARY VARIABLES (from Census/Admin)

- **auxiliary_mean_income** [Numeric]: Average income from tax records
- **auxiliary_unemployment** [Numeric]: Unemployment rate from labor office
- **auxiliary_education_years** [Numeric]: Mean years of schooling
- **auxiliary_asset_index** [Numeric]: Composite asset score
- **auxiliary_dependency_ratio** [Numeric]: Dependent/working age ratio
- **auxiliary_urbanization** [Numeric]: Proportion urban (0-1)

SATELLITE INDICATORS

- **satellite_nightlights** [Numeric]: Nighttime lights intensity
- **satellite_vegetation** [Numeric]: NDVI vegetation index
- **distance_to_road_km** [Numeric]: Distance to nearest paved road
- **distance_to_market_km** [Numeric]: Distance to nearest market

CLIMATE VARIABLES

- **climate_rainfall_mm** [Numeric]: Annual rainfall (millimeters)
- **climate_temperature_c** [Numeric]: Mean annual temperature (Celsius)

MODEL-BASED ESTIMATES

- **fh_estimate** [Numeric]: Fay-Herriot poverty estimate
- **fh_mse** [Numeric]: Mean squared error of FH estimate
- **model_based_estimate** [Numeric]: Alternative model estimate
- **model_based_mse** [Numeric]: MSE of model-based estimate
- **synthetic_estimate** [Numeric]: Regression synthetic estimate
- **composite_estimate** [Numeric]: Weighted composite estimate

QUALITY ASSESSMENT

- **reliability_flag** [String]: Estimate reliability
  - Reliable: CV < 20%
  - Use with caution: CV 20-30%
  - Unreliable: CV > 30%

FAY-HERRIOT MODEL SPECIFICATION

Area-level model:
$y_i = X_i'\beta + v_i + e_i$

where:
- y_i: Direct estimate for area i
- X_i: Vector of auxiliary variables
- v_i: Area random effect ~ $N(0, \sigma_v^2)$
- e_i: Sampling error ~ $N(0, \psi_i)$
- ψ_i: Known sampling variance

AUXILIARY DATA SOURCES

1. **Census 2022**: Demographics, education, housing
2. **Administrative Records**:
   - Tax authority (income)
   - Labor ministry (unemployment)
   - Education ministry (enrollment)
3. **Satellite Data**:
   - VIIRS nighttime lights (2023)
   - MODIS vegetation indices (2023)
4. **Geographic Data**:
   - OpenStreetMap (roads, facilities)
   - Climate databases (WorldClim)

MODEL DIAGNOSTICS

- **Residual Analysis**: Standardized residuals checked
- **Influence Diagnostics**: Leverage points identified
- **Model Fit**: R^2 and AIC compared across models
- **Cross-Validation**: Leave-one-out validation performed

ESTIMATION QUALITY METRICS

Indicator	Target	Acceptable
CV (National)	< 5%	< 10%
CV (District)	< 15%	< 20%
CV (Small Area)	< 25%	< 35%
Coverage Rate	95%	90%
Bias	< 2%	< 5%

BENCHMARKING PROCEDURES

1. **Internal Consistency**: Small areas sum to district totals
2. **External Validation**: Comparison with admin records
3. **Time Consistency**: Reasonable trends from previous estimates

MISSING DATA HANDLING

- **Auxiliary Variables**: Multiple imputation for missing values
- **Direct Estimates**: Areas with sample < 10 use synthetic only
- **Quality Flags**: Applied based on data availability

USAGE GUIDELINES

1. **Poverty Mapping**: Use fh_estimate for best precision
2. **Uncertainty**: Always report with MSE/confidence intervals
3. **Aggregation**: Use proper weighting when combining areas
4. **Time Series**: Account for model changes between years
5. **Policy Use**: Check reliability_flag before decisions

LIMITATIONS

1. **Model Assumptions**: Normality, linear relationships
2. **Auxiliary Quality**: Admin data may have coverage gaps
3. **Spatial Correlation**: Not explicitly modeled
4. **Temporal Lag**: Auxiliary data from different periods

VALIDATION RESULTS

- **Internal Validation R^2**: 0.75

- **External Validation**: Correlation 0.82 with admin data
- **Bias Assessment**: Average bias -1.3%
- **Coverage Properties**: 94% of 95% CIs contain true values

FILE SPECIFICATIONS

- **Format**: CSV
- **Records**: 500 small areas
- **Variables**: 28
- **File Size**: 350 KB
- **Encoding**: UTF-8

VERSION INFORMATION

- **Version**: 1.0
- **Production Date**: 2024-02-15
- **Methodology Document**: SAE_Methodology_v2.pdf
- **Software Used**: R (sae package v1.3)

CONTACT

- **Technical Lead**: Dr. James Mwangi
- **Email**: small.area@sadc-stats.org
- **Phone**: +267 XXX XXXX

CITATION

SADC Statistical Unit (2024). Small Area Poverty Indicators 2024.
Model-based estimates using Fay-Herriot methodology.
Southern African Development Community, Gaborone, Botswana.